

Theoretical Results on the Use of Dimensional and Similarity Analyses in the Statistical Analysis of Computer Codes

José Mira(1), Ricardo Bolado(2) and María Jesús Sánchez (1)

(corresponding author: José Mira)

(1) Laboratorio de Estadística

E. T. S. de Ingenieros Industriales

c/ José Gutierrez Abascal, 2

28006 Madrid (Spain)

Phone: 34913363148

Fax: 34913363149

e-mail: jmira@etsii.upm.es and mjsan@etsii.upm.es

(2) Nuclear Safety Unit, Institute for Energy, Joint Research Center

Westerduinweg 3. NL-1755 Petten. The Netherlands

email: ricardo.bolado-lavin@jrc.nl

June 16, 2003

The purpose of this paper is to present some theoretical results concerning the application of dimensional and similarity analyses to two problems of statistical analysis of computer code outputs (SACCO): first, the propagation of uncertainties and second the prediction of the code output through kriging models.

Since real experimentation is very expensive, with the introduction of faster and cheaper computers, computer code simulation has become an essential research tool of

science and engineering. The relationship between input and output is very frequently expressed through functional equations (differential equations or finite difference equations).

Let us consider a system of functional equations where $\mathbf{Y} = (Y_1, \dots, Y_n)$ are the dependent or output variables and $\mathbf{X} = (X_1, \dots, X_m)$ are the independent variables (e.g., space coordinates and time). Let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$ be the *parameters* of the system, i.e., coefficients of the differential equations and of the initial and boundary conditions. The solutions to the system are $Y_j = I_j(\mathbf{X}, \boldsymbol{\theta})$. Let us assume that the values of some of the $\mathbf{X}, \boldsymbol{\theta}$ are not known precisely, i.e., there is uncertainty about their values, this uncertainty is described through a known probability distribution. This implies that \mathbf{Y} becomes stochastic, and the problem arises of calculating its joint distribution for given values of the non-stochastic subset of \mathbf{X} . In computer code simulation literature, this problem of transformation of variables in probability theory is often called *propagation of uncertainties or uncertainty analysis*. In general, the problem can not be solved analytically, so approximate methods are applied, preferably Monte Carlo, where variance reduction methods are applied to reduce computational costs. The most frequently used variance reduction methods are stratified sampling, importance sampling and Latin Hypercube Sampling (McKay, Conover and Beckman (1979))

Another problem related to the computer simulation of physical processes is the application of the kriging models of geostatistics to the prediction of deterministic functions, particularly the computer codes which simulate the processes by solving, in many cases, a system of differential equations in accordance with the notation defined above. The pioneering paper is Currin et al (1988), Handcock and Stein (1993) present a very useful kriging model for univariate responses and Mira and Sánchez (2002) is an extension to bivariate responses.

Let us now introduce the problem of dimensional and similarity analyses. In physics, one speaks of similarity between two problems when one can transform one problem into the other by a change of scale in the variables, see for instance Langhaar (1951) for a classical reference. It is shown that this is possible when a set of dimensionless numbers (in mathematical terms, we shall speak instead of invariant functions), which are functions of the parameters $\boldsymbol{\theta}$, coincide in both problems. A classical example is the Reynolds number in fluid mechanics. The dimension of the parameter space, originally p , can thus be

reduced to the number of dimensionless quantities which define the system of functional equations as far as the parameters are concerned. An a priori different problem is the reduction of the number of independent variables \mathbf{X} . For example, a partial differential equation can be thus reduced to an ordinary differential equation.

Moran (1971) and Moran and Marshek (1972) have shown that these two problems of reduction of dimensionality, i.e., the reduction in the number of parameters through dimensional analysis and the reduction of the number of independent variables, can be formulated in terms of a more general framework called generalized dimensional analysis which includes both.

Moran and Marshek's generalized dimensional analysis consists in finding a set of linear transformations:

$$\begin{aligned} Y'_j &= K_j Y_j \quad (j = 1, \dots, n \geq 1) \\ X'_k &= K_{n+k} X_k \quad (k = 1, \dots, m \geq 1) \\ \theta'_e &= K_{n+m+e} \theta_e \quad (e = 1, \dots, p \geq 1) \end{aligned} \tag{1}$$

of the $\mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}$, where the K_j , $j = 1, \dots, n + m + p$ are constants, such that the system of functional equations is invariant under the transformations i.e., $Y_j = I_j(\mathbf{X}; \boldsymbol{\theta})$ transforms to $Y'_j = I_j(\mathbf{X}'; \boldsymbol{\theta}')$, where $\mathbf{X}' = (X'_1, \dots, X'_m)$ and $\boldsymbol{\theta}' = (\theta'_1, \dots, \theta'_p)$. We note that the prime symbol stands for variable transformation and not for array transposition. A more general class of transformations could have been used, but we are restricted here to linear transformations (scale changes) because they have proved useful in many physical problems, while maintaining mathematical simplicity and a clear physical interpretation.

When imposing invariance in the case of linear transformations there appear restrictions linking the values of the K_i , $i = 1, \dots, n + m + p$. In most cases, the restrictions will reduce their degrees of freedom so if initially there are $n + m + p$ transformation constants K_i and q restrictions, there will finally be $r = n + m + p - q$ degrees of freedom for the K_i .

We can then define the transformations in terms of a subset of r constants which we call A_j , $j = 1, \dots, r$ and the set of transformations can be rewritten:

$$\begin{aligned}
Y'_j &= A_1^{a_{j1}} \dots A_r^{a_{jr}} Y_j \quad (j = 1, \dots, n \geq 1) \\
X'_k &= A_1^{b_{k1}} \dots A_r^{b_{kr}} X_k \quad (k = 1, \dots, m \geq 1) \\
\theta'_e &= A_1^{c_{e1}} \dots A_r^{c_{er}} \theta_e \quad (e = 1, \dots, p \geq 1)
\end{aligned} \tag{2}$$

where the a_{ji} , b_{kl} and c_{et} are exponents.

Each restriction will define an invariant function

$$\pi = Y_1^{\alpha_1} \dots Y_n^{\alpha_n} X_1^{\beta_1} \dots X_m^{\beta_m} \theta_1^{\gamma_1} \dots \theta_p^{\gamma_p} \tag{3}$$

where the $\alpha_i, \beta_j, \gamma_k$ are also exponents, in such a way that (see Moran and Marsheck (1972)) the system of functional equations can be expressed in terms of these invariant functions, instead of in terms of the original and larger set formed by $\mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}$.

The calculation of the invariants and of the expression of the system of functional equations in terms of the invariants is formalized in the theorems of Moran and Marshek (1972), we assume that a set of transformations of the type (2) has been found such that these transformations depend on a number r of essential parameters A_i . Theorem 1 says that then a number of $q = n + m + p - r$ invariant ("dimensionless" in the physics terminology) functions Π_j and π_k of the form (3) can be found such that the functional relationship of the system can be expressed in terms of n relations between the Π_j and π_k of the form :

$$\Pi_j(Y_j; \mathbf{X}; \boldsymbol{\theta}) = F_j(\pi_1(\mathbf{X}; \boldsymbol{\theta}), \dots, \pi_\delta(\mathbf{X}; \boldsymbol{\theta})) \tag{4}$$

where $\delta = m + p - r$, so, since there are now $m + p - r$ independent variables and parameters instead of the original $m + p$, (the number of outputs n does not change and there are $q = \delta + n$ restrictions) and there is a dimension reduction which can be used for a better analysis of the system of functional equations.

Given the reduction of dimensionality from $m + p$ to $m + p - r$, the invariants π_i which are in general a function of the independent variables \mathbf{X} and the parameters $\boldsymbol{\theta}$ can be built in such way that there is a reduction in either the number of parameters or in both the number of parameters and independent variables. Theorems 2 and 3, respectively, are concerned with the conditions necessary for these situations to appear.

The application of dimensional and similarity analyses to the two above mentioned problems of SACCO consists in : for the propagation of uncertainties, to obtain a reduced space (space of reduced dimension) for the stochastic inputs with no loss of information. Then, variance reduction techniques such as stratified sampling can be applied in this reduced space, instead of sampling in the original space of larger dimension; it is shown that the variance of the estimators of the distribution of the computer code output is reduced while maintaining the sample size. For the kriging predictors of deterministic functions (computer codes) the experimental design is obtained in the reduced input space instead of in the original one, thus improving the properties of the design.

The success of the application of dimensional and similarity analyses to both the propagation of uncertainties and optimal design for kriging predictors of deterministic functions has been shown empirically and justified intuitively in Mira, Bolado and Solana (2003), and in Mira, Bolado, Sánchez and Valero (2002) but not proved as such. A similar application but with a reduction of dimensionality obtained in a less sophisticated way is Bolado and Mira (2003). The computer code used in all cases was a radioactive transport simulation code. The purpose of this paper is to present two theorems in this direction, one for each field of application.

The proof of the theorems is based on the following intuition. Let us consider the application to the propagation of uncertainties, when using stratified sampling on the input space. The key point is to achieve a better stratification of the output space. Stratified sampling consists in placing in the same stratum values which are close so that the strata are as homogeneous as possible. This requires knowing as best as possible, for two given inputs, the distance between the corresponding outputs, and what we are really achieving through dimensional analysis is a better knowledge of these distances: if the distance between the two outputs can be in some way partitioned in two contributions: first, the distance between the values of the invariants, second the distance between the outputs for the invariants, then, through dimensional analysis we can manipulate the first contribution, making it as small as possible in average inside each stratum, and thus reduce the average global distance between the outputs in each stratum.

References

Bolado, R. and Mira, J., (2003), "Trivial Reductions of Dimensionality: An Application to a Radioactive Contaminant Transport Code", *Environmetrics* (in press).

- Currin, C, Mitchel, T., Morris, M., and Ylvisaker, D., (1991), "Bayesian Prediction of Deterministic Functions, with Application to the Design and Analysis of Computer Experiments". *Journal of the American Statistical Association*, vol 86, No. 416.
- Handcock, M., and Stein, M., (1993), "A Bayesian Analysis of Kriging". *Technometrics*, vol 35, No. 3.
- Langhaar, H.L., (1951), *Dimensional Analysis and Theory of Models*. Wiley.
- McKay, Conover and Beckman (1979), "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code". *Technometrics*, vol. 21, No. 2, pp. 239-245.
- Mira, J., Bolado, R. and Solana, P., (2003), "The Use of Dimensional and Similarity Analyses in the Propagation of Uncertainties: A Physical Example", *Journal of Computational and Graphical Statistics*, (in press).
- Mira, J., Bolado, R., Sánchez, M.J., and Valero, J. (2002), "Using Dimensional Analysis for Design Improvement in the Prediction of Deterministic Functions: A Radiative Transport Example"., submitted for publication to *Environmetrics*.
- Mira, J. and Sánchez, M.J., (2002) "Theoretical Results for a Bayesian Bivariate Kriging Model". *Statistics and Probability Letters*, vol. 292, No 6.
- Moran, M., (1971) "A Generalization of Dimensional Analysis". *Journal of the Franklin Institute*, vol. 292, No 6.
- Moran, M., and Marshek, K.M, (1972), "Some Matrix Aspects of Generalized Dimensional Analysis". *Journal of Engineering Mathematics*, Vol. 6, No. 3.